Utah
Academic Language Proficiency Assessment

Technical Report
2006-2007

Copyright © 2007 by the Utah State Office of Education. All rights reserved. No part of this work may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying and recording, or by any information storage or retrieval systems, except as may be expressly permitted in writing by the Utah State Office of Education, 250 East 500 South, P.O. Box 144200, Salt Lake City, Utah 84114-4200.

Utah Academic Language Proficiency Assessment
2006-2007 Technical Report

TABLE OF CONTENTS

	Page
Purpose of the UALPA	05
UALPA Test Design	05
Structure of the UALPA	05
Development of the UALPA	06
UALPA Technical Characteristics	
-Operational Test Administration and Data Analyses	09
-Calibration and Scaling Methodology	10
-Setting Proficiency Standards	11
Reliability of the UALPA	12
Reliability of Handscoring	15
Validity	16
Bias/Sensitivity	17
Summary of the 2006-07 Test Administration	17
Appendix: Report of Equipercentile Linking Study	
-Purpose	18
-Program Details	18
-Technical Considerations	19
-Linking Methods	20
-Results	21
-Comparability Analysis	22

Utah Academic Language Proficiency Assessment
Technical Report
2006-2007

Purpose of the UALPA

The Utah Academic Language Proficiency Assessment (UALPA) is a modified version of an assessment developed by the Mountain West Assessment Consortium to fulfill the requirements of 'No Child Left Behind' (NCLB) legislation. The purpose of the test is to measure the English proficiency of K-12 students who are not native speakers of English and to monitor annual growth in English proficiency. The information provided by the UALPA is intended to support instructional efforts on behalf of English language learners (ELLs) and to help ensure that they attain full English proficiency.

UALPA Test Design

The UALPA assesses English proficiency in Listening, Speaking, Reading, and Writing and reports scores in each of those language modalities as well as in Comprehension (a combination of select items from the Listening and Reading test). In addition, the UALPA reports a total score that reflects overall English proficiency. UALPA test forms were designed for specific grade/grade clusters: K, 1-2, 3-6, 7-8, and 9-12.

Structure of the UALPA

Table 1 shows the structure of the test form for each grade cluster, including the numbers of items (Itm) by item type in each language modality as well as the number of points (Pts) represented by those items. The items and points in the Comprehension column do not contribute to the Totals shown in the last two columns because all Comprehension items are part of the Listening or Reading tests.

Table 1. Structure and Content of UALPA Test Forms

Form	Grade Cluster	Item Type	Listen		Speak		Read		Write		Comp		Total	
			Itm	Pts	Itm	Pts	Itm	Pts	Itm	Pts	Itm	Pts	Itm	Pts
A	K	MC	15	15	-	-	12	12	-	-	20	20	27	27
		SA	5	10	9	18	13	13	6	12	-	-	33	53
		ER	-	-	2	7	-	-	-	-	-	-	2	7
		Total	**20**	**25**	**11**	**25**	**25**	**25**	**6**	**12**	**20**	**20**	**62**	**87**
B	1-2	MC	25	25	-	-	25	25	-	-	35	35	50	50
		SA	-	-	6	12	-	-	19	19	-	-	25	31
		ER	-	-	4	13	-	-	2	6	-	-	6	19
		Total	**25**	**25**	**10**	**25**	**25**	**25**	**21**	**25**	**35**	**35**	**81**	**100**

C	**3-6**	**MC**	25	25	-	-	23	23	6	6	32	32	54	54
		SA	-	-	9	18	-	-	7	7	1	2	16	25
		ER	-	-	2	7	1	2	4	12	-	-	7	21
		Total	**25**	**25**	**11**	**25**	**24**	**25**	**17**	**25**	**33**	**34**	**77**	**100**

D	**7-8**	**MC**	25	25	-	-	23	23	7	7	34	34	55	55
		SA	-	-	9	18			6	6	1	2	15	24
		ER	-	-	2	7	1	2	4	12	-	-	7	21
		Total	**25**	**25**	**11**	**25**	**24**	**25**	**17**	**25**	**35**	**36**	**77**	**100**

E	**9-12**	**MC**	30	30	-	-	26	26	18	18	50	50	74	74
		SA	-	-	10	20	-	-	1	2	-	-	11	22
		ER	-	-	3	10	1	4	3	10	1	4	7	24
		Total	**30**	**30**	**13**	**30**	**27**	**30**	**22**	**30**	**51**	**54**	**92**	**120**

MC - Multiple Choice; SA - Short Answer; ER - Extended Response

Development of the UALPA

The following excerpt from the *Mountain West Consortium Foundation Document* provides background information on the design of the UALPA.

> **Mountain West Assessment Consortium Foundation Document**
> ***Introduction***
>
> The *Mountain West Assessment Consortium Foundation Document* is part of a response to the No Child Left Behind Act (NCLB) of 2001 that mandates assessment of English language learners' progress in attaining proficiency of academic English. Since regular state assessments may not accurately reflect the gains English language learners have made in attaining English proficiency, the Mountain West Assessment Consortium has developed an English language proficiency assessment to serve a dual purpose: to measure students' language proficiency and to measure students' progress toward meeting state standards. Through the development and administration of this assessment, Mountain West Consortium states will satisfy the NCLB requirements for monitoring the development of English proficiency of the English language learners in their public schools.
>
> The *Mountain West Assessment Consortium Foundation Document* describes the elements of language proficiency that are the basis for the Mountain West Assessment Consortium's English Language Proficiency Assessment. The purpose of the assessment is to gauge English language learners' progress in learning to listen to, speak, read, and write in the English language. The assessment follows a developmental progression across and within distinct grade spans. It is based on five communication standards recognized as the linguistic underpinnings of language:

phonology, morphology, vocabulary, syntax, and function. The standards have been further detailed in benchmark performance descriptors.

Standards and benchmark descriptors are common elements of any framework that describes what students should know and be able to do. Standards are like umbrellas; they are broad-based, encompassing a set of related skills and/or knowledge bases. Benchmarks are more specific statements that describe discrete tasks students will perform in order to demonstrate knowledge or skills within a standard. For example, under the vocabulary standard in reading, one benchmark descriptor is, *"Reads and understands common idioms."*

The Mountain West Assessment Consortium English Language Proficiency Assessment includes separate modules for children at these grade spans: kindergarten through early first grade; mid-first grade through second grade; third grade through fifth grade; sixth grade through eighth grade; ninth grade through twelfth grade. Within each of these designated grade spans, assessment items have been developed to evaluate growth in English language acquisition across three broad developmental levels: early acquisition, intermediate, and transitional. The assessment battery modules include test items at each of the three developmental levels across the four modalities of listening, speaking, reading, and writing.

It is important to emphasize the breadth of these developmental levels and to recognize that they are not proficiency levels. The developmental levels of the standards are intentionally broad; they are used simply to make general classifications of test items within the assessment. Proficiency or performance levels specify what a student has achieved or demonstrated *relative to a set of standards*. There may, in fact, be as many as five distinct proficiency levels within these three broad developmental levels. Proficiency or performance levels are determined through standard-setting activities that yield cut-scores within the total range of test scores. There are several ways to determine proficiency levels, and each state that elects to use the Mountain West Assessment Consortium English Language Proficiency Assessment will apply its own process to determine proficiency levels.

Benchmarks have been grouped within five standards to reflect the dimensions of communicative competency:

- Phonology/Orthography standards are used to evaluate students' progress in understanding and correctly manipulating the sound system of English.
- Morphology standards are used to evaluate students' progress in understanding and using the rules of English word formation.

- Vocabulary standards are used to evaluate students' understanding and appropriate use of English words and phrases (semantic knowledge).
- Syntax standards are used to evaluate students' progress in understanding and using the rules of English sentence formation.
- Function/Discourse standards are used to evaluate students' ability to use and comprehend English in various oral and written contexts.

Since elements of some standards must be in place before others develop, the application of these five language standards varies across both grade spans and developmental levels. For example, phonology benchmarks are generally addressed more extensively at the early acquisition level than at intermediate or transitional levels. In addition, the requirements for competency in the four modalities (listening, speaking, reading, and writing) vary so that one modality may emphasize some standards over others. For example, expectations for syntax use are more pronounced in the language production modalities of speaking and writing. Similarly, assessment of function/discourse skills is addressed in greatest depth at the transitional level.

All of the standards and benchmarks included in this document are addressed in the assessment. The majority of the benchmarks are addressed in specific assessment tasks. Other benchmarks are addressed indirectly through holistic acts of listening, speaking, reading, or writing. In the receptive processes of listening or reading, acquisition of some benchmarks is inherent in demonstrations of comprehension of the language presented. Holistic scoring rubrics have been developed to encompass such benchmarks in the language production modalities of speaking and writing.

The order in which progress across the four language modalities is assessed also reflects a developmental perspective. The modalities generally considered informal - listening and speaking - precede assessment of the more formal language modalities of reading and writing. Moreover, since a degree of language comprehension generally precedes language production, receptive language skills are addressed before production skills in both informal and formal order in the assessment. Thus, listening skills are assessed first, followed by speaking, reading, and writing skills in that order.

The developmental continuum is also reflected in this assessment in the degree to which language is decontextualized. At the early acquisition level, care has been taken to provide directions that are simple and concrete. Demonstration and practice items are also provided to help students understand what is expected of them. In addition, language in the test directions for intermediate and transitional level items begins to approximate the language found in mainstream assessments.

Additional information and details about the content of the assessment are included in the full text of the *Mountain West Consortium Foundation Document*.

UALPA Technical Characteristics

Operational Test Administration and Data Analyses

The UALPA was administered during the 2006-2007 academic year. Handscoring of open-ended writing items and scanning of answer documents were completed by Measured Progress. The details of those operations were not available for this report. At the end of the academic year, a scored, scan file was made available to Questar Assessment. The summary data presented in this report and the technical work completed by Questar were based on that file.

Item analyses included calculation of item difficulty and item discrimination. The typical classical item difficulty measure is the p-value statistic which is calculated as the proportion of students that responded correctly for each MC item or the average percent of the maximum score that students earned on each CR item. Thus p-value is inversely related to item difficulty. Generally, p-values should range between 0.30 and 0.90. Table 2 shows the average p-value and range for items by language modality and test form. Several generalizations can be made the difficulty of UALPA items. Generally item difficulty fell within the accepted range but consistently exceeded 0.90. Based on average difficulty, Speaking tests were the easiest and Reading or Writing tests were the most difficult.

Point-biserial correlations are used to examine item discrimination. Point-biserial correlation coefficients range between -1.0 and +1.0 and indicate the extent to which an item differentiates between low-ability and high-ability students. High positive values indicate that a high-ability student is more likely to answer an item correctly and low negative values indicate that a low-ability student is more likely to answer an item correctly. Table 2 shows median point-biserial correlation coefficients and range by language modality and test form. Typically, point-biserial correlations range between 0.3 and 0.6 with values below 0.2 indicating potentially problematic items. The summary statistics in Table 2 show that although median point-biserial coefficients fell within the accepted range, there were some values below 0.20.

Table 2. Summary of Item Difficulty and Discrimination by Grade Cluster and Language Modality.

Grade Span	Modality	Item p-value		Point Biserial	
		Average	Range	Median	Range
K	L	0.67	0.37 - 0.95	0.38	0.26 - 0.62
	S	0.67	0.40 - 0.88	0.58	0.51 - 0.62
	R	0.56	0.31 - 0.76	0.46	0.10 - 0.61
	W	0.66	0.39 - 0.94	0.38	0.01 - 0.46
1-2	L	0.78	0.49 - 0.99	0.33	0.18 - 0.47
	S	0.85	0.68 - 0.97	0.46	0.34 - 0.58
	R	0.79	0.37 - 0.97	0.44	0.24 - 0.55
	W	0.68	0.30 - 0.95	0.51	0.15 - 0.67
3-6	L	0.86	0.60 - 0.97	0.34	0.25 - 0.47
	S	0.93	0.81 - 0.99	0.49	0.23 - 0.62
	R	0.85	0.36 - 0.98	0.42	0.21 - 0.55
	W	0.78	0.44 - 0.94	0.44	0.24 - 0.62
7-8	L	0.78	0.44 - 0.98	0.33	0.11 - 0.53
	S	0.89	0.79 - 0.96	0.61	0.49 - 0.69
	R	0.72	0.28 - 0.95	0.38	0.04 - 0.54
	W	0.73	0.33 - 0.95	0.46	0.06 - 0.56
9-12	L	0.75	0.48 - 0.98	0.39	0.16 - 0.58
	S	0.87	0.75 - 0.95	0.64	0.49 - 0.71
	R	0.74	0.26 - 0.96	0.39	0.27 - 0.54
	W	0.68	0.28 - 0.93	0.41	0.19 - 0.62

Analyses of test level data, including raw score descriptive statistics and test reliability measures, are reported in Table 4.

Calibration and Scaling Methodology

UALPA items were calibrated using Winsteps (v. 3.57) software implementing the Rasch Partial Credit Model which can accommodate both multiple-choice and open-response items. The data

file obtained from Measured Progress was used for the item calibration. Each grade cluster was calibrated separately. Sample sizes per grade are shown in Table 4.

Separate Raw Score to Scaled Score conversion tables were created for each form. These tables were based on the raw score to Rasch ability (theta) tables produced during the calibration. Scaled scores were produced by a linear transformation of the theta associated with each raw score. Raw score to scaled score conversion tables were created for UALPA Total Test as well as for each language modality (S, L, R, W) and Comprehension.

Setting Proficiency Standards

A formal UALPA Standard Setting was conducted June 26-28, 2007. That process involved 36 Utah educators divided into two panels: One panel focused on grades K through 6 and the other panel focused on grades 7 through 12. A detailed report of standard setting activities and recommendations has been submitted to the USOE. A brief summary of procedures and results follows.

Panel members made recommendations for one grade span at a time. One panel began with grades K-2 and the second panel began with grades 7-8. After reviewing test forms for the grade span under consideration, panelists received ordered item booklets containing test items for the grade span, with each page displaying one test item and pages ordered in terms of increasing item difficulty. Using the Bookmark or item mapping procedure, panelists made “cuts” by placing markers in the books to indicate the item on which 50% of the students at a particular proficiency level and in a particular grade would answer correctly. Panelists were instructed to proceed through the books by grade within the grade span. Three rounds of cuts were planned for each grade span. Following each of the first two rounds, panelists were shown frequency distributions and medians of recommended cuts and were given the opportunity to discuss the process. The second round was followed by impact data, i.e., the percent of students in each grade who would be placed in each proficiency level based on the median cuts assigned by the group. The third round of cuts was accepted as the panelists’ final recommendations.

The panelists’ final recommendations were adjusted to eliminate minor variations within grade clusters and to create a more consistent pattern of proficiency levels across the grades. These final panelists’ recommendations with adjustments were presented to USOE and subjected to a second set of adjustments prior to being accepted. Panelists’ recommendations as well as each set of adjustments and likely impact data are documented in the UALPA Standard Setting Final Report. Table 3 shows, for each form and grade, the range of UALPA scaled scores corresponding to each proficiency level.

Table 3. Total UALPA Scaled Scores Corresponding to Proficiency Levels

		Total UALPA Proficiency Levels			
Form	**Grade**	**PreEmergent**	**Emergent**	**Intermediate**	**Advanced**
A	**K**	Below 227	227 - 258	259 - 285	286 and above
B	**1**	Below 239	239- 256	257- 303	304 and above
	2	Below 250	250 - 279	280- 337	338 and above
C	**3**	Below 239	239 - 275	276 - 331	332 and above
	4	Below 239	339 - 2772	278 - 345	346 and above
	5	Below 248	248 - 280	281 - 358	359 and above
	6	Below 248	248 - 288	289 - 358	359 and above
D	**7**	Below 213	213 - 254	255 - 316	317 and above
	8	Below 213	213 - 254	255 - 319	320 and above
E	**9**	Below 213	213 - 257	258 - 309	310 and above
	10	Below 213	213 - 258	259 - 314	315 and above
	11	Below 219	219 - 257	258 - 314	315 and above
	12	Below 219	219 - 256	257 - 314	315 and above

Reliability of the UALPA

Data bearing on the reliability of UALPA-1 Test Forms are shown in Table 4. The tables show, for each form and each language modality (and comprehension and the total test), the number of students (N) who were administered the form, coefficient Alpha, a measure of internal-consistency reliability, the maximum raw score attainable, and the mean, standard deviation, and standard error of measurement (SEM) in raw score units. In addition, mean, standard deviation, and SEM are presented for the total test in scaled score units. (Note: N represents the number for whom there was a valid test score and may vary across language domains in a grade to the extent that there were students who did not attempt one or more of the language domain tests. There is a total score for each student regardless of whether or not all language domain tests were attempted.)

Table 4. Reliability, Raw Score and Scaled Score Descriptive Statistics for UALPA Test Forms by Grade

Grade K				**Raw Scores**				**Scaled Scores**		
Form	**Language Modality**	**N**	**Alpha**	**Max**	**Mean**	**Std. Dev.**	**SEM**	**Mean**	**Std. Dev.**	**SEM**
A (K)	Listening	4,782	0.86	25	16.0	5.81	2.20	-	-	-
	Speaking	4,782	0.89	25	15.9	6.47	2.13	-	-	-
	Reading	4,782	0.90	25	13.7	6.60	2.08	-	-	-
	Writing	4,782	0.72	12	7.5	3.06	1.62	-	-	-
	Comprehen	4,782	0.86	20	16.0	5.81	2.20	-	-	-
	Total	4,782	0.94	87	53.1	17.91	4.35	267.6	34.58	8.40

Grade 1

				Raw Scores				Scaled Scores		
Form	**Language Modality**	**N**	**Alpha**	**Max**	**Mean**	**Std. Dev.**	**SEM**	**Mean**	**Std. Dev.**	**SEM**
B (1-2)	Listening	4,855	0.83	25	17.8	4.65	1.90	-	-	-
	Speaking	4,855	0.88	25	19.2	5.78	2.04	-	-	-
	Reading	4,855	0.88	25	17.7	5.32	1.81	-	-	-
	Writing	4,855	0.89	25	13.0	5.60	1.89	-	-	-
	Comprehen	4,855	0.87	35	24.0	6.44	2.30	-	-	-
	Total	4,855	0.95	100	67.7	18.16	4.10	283.5	43.86	9.91

Grade 2

Form	Language Modality	N	Alpha	Max	Mean	Std. Dev.	SEM	Mean	Std. Dev.	SEM
B (1-2)	Listening	4,725	0.87	25	20.4	4.49	1.64	-	-	-
	Speaking	4,725	0.90	25	21.3	5.26	1.69	-	-	-
	Reading	4,725	0.91	25	21.0	4.73	1.44	-	-	-
	Writing	4,725	0.89	25	17.7	5.35	1.80	-	-	-
	Comprehen	4,725	0.90	35	28.3	6.16	1.94	-		-
	Total	4,725	0.96	100	80.4	17.44	3.53	315.2	50.25	10.18

Grade 3

				Raw Scores				Scaled Scores		
Form	**Language Modality**	**N**	**Alpha**	**Max**	**Mean**	**Std. Dev.**	**SEM**	**Mean**	**Std. Dev.**	**SEM**
C (3-6)	Listening	4,240	0.88	25	20.3	4.73	1.63	-	-	-
	Speaking	4,240	0.89	25	22.2	4.73	1.57	-	-	-
	Reading	4,240	0.90	25	19.4	4.84	1.56	-	-	-
	Writing	4,240	0.87	25	15.9	5.20	1.87	-	-	-
	Comprehen	4,240	0.90	34	26.2	6.20	1.95	-	-	-
	Total	4,240	0.96	100	77.8	17.30	3.50	307.4	47.25	9.57

Grade 4

Form	Language Modality	N	Alpha	Max	Mean	Std. Dev.	SEM	Mean	Std. Dev.	SEM
C (3-6)	Listening	3,407	0.91	25	20.9	5.01	1.49	-	-	-
	Speaking	3,407	0.92	25	22.3	5.20	1.50	-	-	-
	Reading	3,407	0.92	25	20.1	5.03	1.45	-	-	-
	Writing	3,407	0.89	25	17.2	5.44	1.82	-	-	-
	Comprehen	3,407	0.93	34	27.3	6.65	1.80	-	-	-
	Total	3,407	0.97	100	80.5	18.83	3.31	316.9	54.15	9.53

Grade 5

Form	Language Modality	N	Alpha	Raw Scores Max	Raw Scores Mean	Raw Scores Std. Dev.	Raw Scores SEM	Scaled Scores Mean	Scaled Scores Std. Dev.	Scaled Scores SEM
C (3-6)	Listening	3,238	0.91	25	21.5	4.73	1.40	-	-	-
	Speaking	3,238	0.91	25	22.6	4.90	1.45	-	-	-
	Reading	3,238	0.91	25	20.9	4.62	1.38	-	-	-
	Writing	3,238	0.88	25	18.3	5.14	1.78	-	-	-
	Comprehen	3,238	0.93	34	28.3	6.24	1.70	-	-	-
	Total	3,238	0.97	100	83.2	17.67	3.21	326.5	51.96	9.44

Grade 6

Form	Language Modality	N	Alpha	Max	Mean	Std. Dev.	SEM	Mean	Std. Dev.	SEM
C (3-6)	Listening	2,691	0.91	25	21.8	4.46	1.36	-	-	-
	Speaking	2,691	0.92	25	22.6	5.02	1.40	-	-	-
	Reading	2,691	0.91	25	21.1	4.47	1.34	-	-	-
	Writing	2,691	0.88	25	18.9	5.06	1.76	-	-	-
	Comprehen	2,691	0.92	34	28.7	5.93	1.65	-	-	-
	Total	2,691	0.97	100	84.5	17.35	3.15	332.2	51.67	9.39

Grade 7

Form	Language Modality	N	Alpha	Raw Scores Max	Raw Scores Mean	Raw Scores Std. Dev.	Raw Scores SEM	Scaled Scores Mean	Scaled Scores Std. Dev.	Scaled Scores SEM
D (7-8)	Listening	2,146	0.85	25	19.0	4.61	1.80	-	-	-
	Speaking	2,146	0.94	25	21.7	5.85	1.42	-	-	-
	Reading	2,146	0.85	25	17.2	4.91	1.88	-	-	-
	Writing	2,146	0.85	25	16.1	5.00	1.94	-	-	-
	Comprehen	2,146	0.88	36	24.8	6.58	2.31	-	-	-
	Total	2,146	0.95	100	74.0	17.49	3.79	294.3	37.97	8.23

Grade 08

Form	Language Modality	N	Alpha	Max	Mean	Std. Dev.	SEM	Mean	Std. Dev.	SEM
D (7-8)	Listening	1,808	0.85	25	19.1	4.60	1.79	-	-	-
	Speaking	1,808	0.94	25	21.6	5.94	1.47	-	-	-
	Reading	1,808	0.86	25	17.6	5.01	1.86	-	-	-
	Writing	1,808	0.86	25	16.6	5.21	1.98	-	-	-
	Comprehen	1,808	0.88	36	25.3	6.66	2.29	-	-	-
	Total	1,808	0.95	100	74.8	17.93	3.85	297.2	39.44	8.46

Grade 09

Form	Language Modality	N	Alpha	Raw Scores Max	Raw Scores Mean	Raw Scores Std. Dev.	Raw Scores SEM	Scaled Scores Mean	Scaled Scores Std. Dev.	Scaled Scores SEM
E (9-12)	Listening	1,544	0.89	30	21.9	6.13	2.06	-	-	-
	Speaking	1,544	0.95	30	25.2	7.60	1.75	-	-	-
	Reading	1,544	0.89	30	19.8	6.10	2.04	-	-	-
	Writing	1,544	0.86	30	18.2	6.20	2.29	-	-	-
	Comprehen	1,544	0.93	54	36.6	10.32	2.83	-	-	-
	Total	1,544	0.96	120	85.0	22.03	4.46	287.9	35.85	7.26

Grade 10

Form	Language Modality	N	Alpha	Max	Mean	Std. Dev.	SEM	Mean	Std. Dev.	SEM
E (9-12)	Listening	1,298	0.90	30	21.6	6.60	2.06	-	-	-
	Speaking	1,298	0.95	30	24.7	8.14	1.86	-	-	-
	Reading	1,298	0.90	30	19.9	6.46	2.03	-	-	-
	Writing	1,298	0.88	30	18.2	6.61	2.33	-	-	-
	Comprehen	1,298	0.94	54	36.5	11.17	2.83	-	-	-
	Total	1,298	0.96	120	84.4	23.68	4.55	287.8	38.41	7.39

Grade 11

Form	Language Modality	N	Alpha	Max	Mean	Std. Dev.	SEM	Mean	Std. Dev.	SEM
E (9-12)	Listening	945	0.90	30	21.8	6.52	2.04	-	-	-
	Speaking	945	0.95	30	25.2	7.76	1.79	-	-	-
	Reading	945	0.90	30	20.4	6.29	2.02	-	-	-
	Writing	945	0.88	30	18.9	6.52	2.30	-	-	-
	Comprehen	945	0.93	54	37.2	10.86	2.81	-	-	-
	Total	945	0.96	120	86.3	22.46	4.49	290.7	38.98	7.80

Grade 12

Form	Language Modality	N	Alpha	Max	Mean	Std. Dev.	SEM	Mean	Std. Dev.	SEM
E (9-12)	Listening	603	0.93	30	21.5	7.36	2.00	-	-	-
	Speaking	603	0.96	30	24.7	8.50	1.72	-	-	-
	Reading	603	0.92	30	19.9	7.03	2.00	-	-	-
	Writing	603	0.90	30	18.2	7.30	2.31	-	-	-
	Comprehen	603	0.95	54	36.5	12.08	2.78	-	-	-
	Total	603	0.97	120	84.3	25.25	4.52	287.1	44.87	8.03

Reliability of Handscoring

Information bearing on the reliability of handscoring is not available for the 2006-07 administration of UALPA.

Validity

Validity refers to the extent to which evidence and theory support the interpretations and inferences made from the scores reported for an assessment. Thus, validation is an ongoing process of accumulating evidence, some of that evidence deriving from analyses of test content and some deriving from analyses of test results.

The structure within which the UALPA was developed and the process of developing test content, as detailed in the Mountain West Assessment Consortium Foundation Document and the Mountain West Assessment Consortium: Final Report, bear on the issue of validity. Several important elements of that process included the development and adoption of a set of common English language development (ELD) standards and a set of learning activities, the development of a Test Blueprint and test design, and item development. Practitioners from individual states were involved throughout the process, receiving support at key points in the process from nationally recognized experts. Local specialists and educators, selected on the basis of a set of relevant criteria, wrote items. More complete details of the content development can be found in the MWAC Final Report.

Reliability is a necessary criterion for validity. Data presented in Table 4 indicate that coefficient alpha, a measure of internal consistency, is above 0.90 for the UALPA Total Test in every grade assessed and, with the exception of Writing in Kindergarten, alphas for every language modality subtest are above 0.80. In addition, the standard error of measurement, which affects the accuracy with which inferences can be made, is acceptably low.

Additional evidence related to the internal structure of the UALPA is presented in Table 5. This table shows intercorrelations among the language modality test scores (**L**istening, **S**peaking, **R**eading, **W**riting, and **C**omprehension) within each grade cluster. The number below the correlation coefficient in each cell represents the number of paired scores on which the correlation is based. All correlation coefficients in this table are significantly different from zero.

Table 5. Intercorrelations of Language Modality Test Scores by Grade Cluster

Grade Cluster	L x S	L x R	L x W	S x R	S x W	S x C	R x W	W x C	Avg.
K	0.67 4,782	0.64 4,782	0.37 4,782	0.56 4,782	0.37 4,782	0.671 4,782	0.41 4,782	0.37 4,782	0.51
1-2	0.64 9,580	0.75 9,580	0.72 9,580	0.61 9,580	0.59 9,580	0.66 9,580	0.82 9,580	0.80 9,580	0.70
3-6	0.74 13,576	0.81 13,576	0.76 13,576	0.70 13,576	0.69 13,576	0.75 13,576	0.83 13,576	0.83 13,576	0.76
7-8	0.57 3,954	0.74 3,954	0.73 3,954	0.55 3,954	0.60 3,954	0.58 3,954	0.78 3,954	0.79 3,954	0.67
9-12	0.52 4,390	0.75 4,390	0.70 4,390	0.48 4,390	0.51 4,390	0.53 4,390	0.79 4,390	0.79 4,390	0.63
Avg.	0.63	0.74	0.66	0.58	0.55	0.64	0.73	0.72	

An equipercentile linking study, conducted by Measured Progress and reported in the Appendix, is also relevant to establishing the validity of UALPA. Although the IPT and the UALPA were administered in different years, the extent to which the distributions of scores is similar suggests that the two instruments are measuring similar constructs.

Going forward, additional evidence bearing on the validity of inferences made on the basis of UALPA results will be gathered.

Bias/Sensitivity

Minimizing item bias is important in several related respects. First, minimizing item bias reduces construct irrelevant variance thus improving validity. Second, minimizing item bias is necessary in establishing the fairness of the instrument for students from diverse cultural and linguistic backgrounds. In addition to the absence of bias, however, the instrument should reflect sensitivity to this diversity. Several different elements of test development are relevant to the issue of bias. The criteria on which item writers were selected, emphasizing credentials and/or experience working with English language learners, provide one essential step in minimizing bias in test content. In addition, prior to their inclusion in test forms, all stimuli, graphics, and items were presented to state selected participants for bias and sensitivity review. Although these procedures are important elements of producing an assessment that minimizes bias, it is also important to utilize empirical procedures for investigating potential bias. Although no such procedures were conducted during the 2006-2007 testing cycle, statistical procedures for investigating differential item functioning (DIF) will be implemented in the current testing year, 2007-2008.

Summary of the 2006-07 Test Administration

[This section will address issues in test administration, scoring, reporting for the year.]

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999). *Standards for educational and psychological testing.*

Measured Progress & Utah State Office of Education. (2005). *Mountain West Assessment Consortium: Final Report.*

Appendix: Report of Equipercentile Linking Study Submitted by Measured Progress

UALPA Technical Summary

Purpose

The purpose of this document is to summarize the statistical links that were made between the IPT assessment program and the new UALPA program. In particular a linking method was used to depict growth occurring between the IPT and UALPA programs. This was needed because the IPT was last administered in 2005/2006 while the UALPA program was first administered in the 2006/2007 school year.

Program Details

The UALPA assessment is designed to measure academic language proficiency in listening, speaking, reading, and writing. The assessment materials are developed based on specific grade spans. The UALPA grade spans are Kindergarten, 1-2, 3-6, 7-8, and 9-12.

Tables 1 and 2 provide an overview of the UALPA test materials and test administration formats. The UALPA test materials are designed to be age appropriate and meet the administration format requirements. For example, in Kindergarten, the assessment is presented as a flip chart with a separate answer document to be completed by the test administrator. In grade span 1-2, the student responses are collected in a consumable test booklet.

Table 1: UALPA Test Materials

Grade Span	Student Booklet Format	Test Admin Manual (w/item content)	Modalities
K	Flipchart	Included in flipchart	Listening, Speaking, Reading, Writing (checklist)
1-2	Consumable scannable	Separate	Listening, Reading, Writing
3-6	Student Booklet	Separate	Listening, Reading, Writing
7-8	Student Booklet	Separate	Listening, Reading, Writing
9-12	Student Booklet	Separate	Listening, Reading, Writing
Speaking A (1-2, 3-6)	Flipchart	Included in flipchart	Speaking
Speaking B (7-8, 9-12)	Flipchart	Included in flipchart	Speaking
Answer Docs (K, 1-2, 3-6, 7-8, 9-12)	Scannable –all modalities in one document	N/A	Listening, Speaking, Reading, Writing

Table 2: UALPA Test Administration Format

Grade Span	Speaking (scored by examiner)	Listening	Reading	Writing
K	Individually administered			
1-2	Individual	Group	Group	Group
3-6	Individual	Group	Group	Group
7-8	Individual	Group	Group	Group
9-12	Individual	Group	Group	Group

The UALPA program differs from the IPT assessment in several ways. The following is a brief overview of the differences between the two programs.

- IPT is designed to measure social language proficiency, rather than academic language.
- Various forms of the IPT are used throughout the state of Utah, where UALPA only has one form that was administered in 2006/2007.
- The grade spans for IPT are 2-3, 4-6, and 7-12. IPT does not test Kindergarten or grade 1.

The difference in construct and other variables between the two programs supported the decision to use an equipercentile linking method to show growth between 2005/2006 and 2006/2007. The decision to do equipercentile linking between the two programs and assessment years was presented to and approved by the Utah Technical Advisory Committee (TAC).

Technical Considerations

IPT Analysis. Using the USOE datasets provided to Measured Progress, a team of psychometricians and data analysts determined the number and percentages of students in each proficiency category of the IPT program. This served as the basis for determining the appropriate cutscores on the UALPA observed student distributions. This preliminary analysis was presented to the USOE to ensure that the percentages of students in each proficiency level matched the state's analysis of the ELL population. Preliminary results were also presented to the TAC, as requested by the USOE.

For the IPT program it is important to note that the grade spans are somewhat different that what is used in the UALPA program. In order to work around this, the analysis that was performed was done per each grade level of the IPT and then weighted results were used against the UALPA observed datasets.

The results of the IPT analysis can be seen in Table 3 where we present the number and percentage of students in each of the IPT proficiency categories (designated as A through D). For example, in the grade span 1-2 we see that there were 1938 students were categorized in Level A, while 9154 students were categorized in Level B. As discussed above, the IPT analysis was done for each grade level and aggregated according to the grade spans of the UALPA program.

Table 3: Percentage of Students at Each IPT Level (2005-2006 IPT Data)

Grade Span	IPT Level	Number of Students	Percentage	Cumulative Percentage
K	A	2041	29.54118	29.54118
	B	3920	56.73759	86.27877
	C	151	2.185555	88.46432
	D	797	11.53568	100
	All	6909	100	
1-2	A	1938	12.74748	12.74748
	B	9154	60.2118	72.95928
	C	1781	11.71479	84.67408
	D	2330	15.32592	100
	All	15203	100	
3-6	A	1643	6.439349	6.439349
	B	9391	36.8058	43.24515
	C	5746	22.52009	65.76524
	D	8735	34.23476	100
	All	25515	100	
7-8	A	571	5.429305	5.429305
	B	3248	30.88333	36.31264
	C	1770	16.82989	53.14253
	D	4928	46.85747	100
	All	10517	100	
9-12	A	1029	5.502674	5.502674
	B	5535	29.59893	35.1016
	C	2513	13.4385	48.54011
	D	9623	51.45989	100
	All	18700	100	

Linking Methods

The method used to map the proficiency distribution results of the IPT onto the UALPA scale was the equipercentile method. That is, by using the distributional results from the IPT program, and the cumulative density functions from the UALPA program, new cutscores on the UALPA raw score scale were found that created similar proficiency distributions for the UALPA program.

Although the method employed here is not strictly speaking an "equating method" per se, it reasonable to use this approach as an attempt to create cutscores between two diverse assessment programs in an attempt to show growth that may have occurred over the two years where IPT and UALPA were administered.

It is also important to note that the IPT does not include the testing of Kindergarten or grade 1 students. To include both grades in the analysis and present cutscores for these grades, psychometricians used a basic statistical extrapolation procedures where IPT data from grades 2, 3 and 4 were used to estimate the percentage of students that might be categorized in the various

proficiency levels for Kindergarten and grade 1 students had they been administered the IPT assessment.

Results

The primary results of the analysis are provided in Table 4, where we present the cutscores found using the above procedures. Specifically, the raw score ranges for each proficiency level have been established from the frequency distributions observed from the administration of the UALPA. The resulting percentage are very similar to those presented in Table 1, the differences are due to the fact that it is necessary to use observed raw scores and specific percentage values cannot be obtained.

For each proficiency category we have provided the range of raw scores points. For example, in grade span 7-8 we see a raw score range of 0 to 35 associated with the A proficiency level (and this will have approximately 5 ½ percent of the students in this category). Also for this grade span we see that over 46% will be categorized at the D level.

Table 4: UALPA Raw Score Range Corresponding to the Percentage of Students at Each IPT Level (2006-2007 UALPA Data)

Grade Span	IPT Level	UALPA Raw Score Range	Number of Students	Percentage	Cumulative Percentage
K	A	0-44	1391	28.9973	28.99729
	B	45-72	2737	57.0565	86.05378
	C	73	81	1.68856	87.74234
	D	74-93	588	12.25766	100
	All	All	4797	100	
1-2	A	0-53	1197	12.461	12.46096
	B	54-86	5704	59.3796	71.84052
	C	87-90	1228	12.7837	84.62419
	D	91-110	1477	15.37581	100
	All	All	9606	100	
3-6	A	0-51	870	6.40082	6.400824
	B	52-85	5199	38.2504	44.65127
	C	86-89	2645	19.46	64.11124
	D	90-109	4878	35.88876	100
	All	All	13596	100	
7-8	A	0-35	220	5.54995	5.54995
	B	36-75	1218	30.7265	36.27649
	C	76-80	669	16.8769	53.15338
	D	81-102	1857	46.84662	100
	All	All	3964	100	
9-12	A	0-35	246	5.59472	5.594724
	B	36-84	1313	29.8613	35.45599
	C	85-91	580	13.1908	48.6468
	D	92-118	2258	51.3532	100
	All	All	4397	100	

Comparability Analysis

In order to justify this type of analysis one must demonstrate that the populations taking the IPT program and the UALPA program are comparable. In other words, it is required that both populations have similar demographic (among other) characteristics. In Tables 5 and 6 we have provided the results of this type of analysis.

A cursory review of Table 5 clearly shows that from a gender perspective the populations taking IPT and UALPA are indeed very similar. The largest difference in males and females can be found in the 9-12 grade span where around 45% of the students were female, and 55% were male. Nonetheless, this was found for both the IPT and for UALPA, and is evidence to suggest that the populations are similar in their gender characteristics.

Table 5: Gender Distributions

Grade Span	Gender	IPT Data		UALPA Data	
		Number of Students	Percentage	Number of Students	Percentage
K	Female	3411	48.7913	2282	48.04211
	Male	3580	51.2087	2468	51.95789
	All	6991	100	4750	100
1-2	Female	7524	48.22459	4735	49.47236
	Male	8078	51.77541	4836	50.52764
	All	15602	100	9571	100
3-6	Female	13830	48.64922	6239	46.10553
	Male	14598	51.35078	7293	53.89447
	All	28428	100	13532	100
7-8	Female	6154	47.49556	1793	45.43842
	Male	6803	52.50444	2153	54.56158
	All	12957	100	3946	100
9-12	Female	10836	45.87249	1963	44.95077
	Male	12786	54.12751	2404	55.04923
	All	23622	100	4367	100

The results for the ethnicity analysis are presented in Table 6. Although there are differences between the IPT and UALPA programs, it is necessary to consider a combination of factors. Two primary factors include sample size, and the nature of the subpopulations involved. For example, it may be reasonable to consider that a given subpopulation may have more migratory tendencies than other subpopulations.

Table 6: Ethnicity Distributions

Grade Span	Ethnicity	IPT Data		UALPA Data	
		Number of Students	Percentage	Number of Students	Percentage
K	A	473	6.765842	205	4.330376
	B	94	1.344586	96	2.027883
	C	447	6.393935	243	5.13308
	H	5677	81.20441	4009	84.68526
	I	74	1.058504	40	0.844951
	P	200	2.860821	123	2.598226
	U	26	0.371907	18	0.380228
	All	6991	100	4734	100
1-2	A	1044	6.69145	553	5.797253
	B	346	2.217664	189	1.98134
	C	1335	8.556595	536	5.619038
	H	11854	75.97744	7812	81.89538
	I	304	1.948468	103	1.079778
	P	622	3.986668	310	3.249817
	U	97	0.621715	36	0.377398
	All	15602	100	9539	100
3-6	A	2076	7.302916	621	4.601022
	B	488	1.716678	298	2.207898
	C	2456	8.639674	743	5.504927
	H	21315	74.98153	11100	82.2405
	I	597	2.100116	181	1.341039
	P	1329	4.675133	492	3.645255
	U	166	0.583952	62	0.459361
	All	28427	100	13497	100
7-8	A	1152	8.890947	182	4.629865
	B	193	1.489542	100	2.543882
	C	1017	7.849039	162	4.121089
	H	9315	71.89164	3250	82.67616
	I	381	2.940495	64	1.628084
	P	816	6.297754	158	4.019334
	U	83	0.64058	15	0.381582
	All	12957	100	3931	100
9-12	A	2188	9.262552	258	5.926947
	B	539	2.281771	152	3.491845
	C	2083	8.818051	224	5.145876
	H	15893	67.2805	3400	78.10705
	I	865	3.661841	114	2.618884
	P	1896	8.026416	172	3.951298
	U	158	0.668868	33	0.758098
	All	23622	100	4353	100